SMART-I^2: "SPATIAL MULTI-USER AUDIO-VISUAL REAL-TIME INTERACTIVE INTERFACE", A BROADCAST APPLICATION CONTEXT.

*Marc Rébillat**, *Brian F.G. Katz*

LIMSI-CNRS
Université de Paris Sud XI
91403 ORSAY, France

Etienne Corteel

sonic emotion
Eichweg 6
CH-8154 OBERGLATT, Switzerland

ABSTRACT

SMART-I^2 is a high quality 3D audio-visual interactive rendering system. In SMART-I^2, the screen is also used as a multichannel loudspeaker. The spatial audio rendering is based on Wave Field Synthesis, an approach that creates a coherent spatial perception of a spatial sound scene over a large listening area. The azimuth localization accuracy of the system has been verified by a perceptual experiment. Contrary to conventional systems, SMART-I^2 is able to realize a high degree of 3D audio-visual integration with almost no compromise on either the audio or the graphics rendering quality. Such a system can provide benefits to a wide range of applications.

Index Terms— Audio-visual integration, wave field synthesis, large multi-actuator panels, virtual reality environments.

1. INTRODUCTION

A great deal of effort has been seen in recent years towards achieving high quality spatialized audio rendering and 3D visual rendering. Despite this, few systems have been conceived as an efficient means of achieving high quality results for both the audio and visual techologies. In any audio-visual (AV) application, the sensation of immersion [1] and the intelligibility of the scenes [2] depend highly on the quality of both the audio and the visual renderings. It is thus important to fulfill the requirements for both technologies in order to achieve a perceptually consistent rendering.

Wave field synthesis [3] (WFS) is a technology which allows one to reproduce a given sound field in a large immersion area. Tracked passive stereoscopy (TPS) is a straight forward means of providing a 3D visual rendering from any viewing position. Both of these technologies place specific demands or constraints on screen and loudspeaker positioning. They also need unobstructed propagation pathways. To meet these constraints, a combination of multi-actuators panels (MAPs) [4] constructed on a large scale, combined with stereoscopic frontal projection has been proposed. This is a technical solution to the design constraints by physically combining the screen and loudspeakers array.

Originally developped for interactive Virtual Reality (VR) applications, this spatialized high quality AV environment can be used as a general AV window into virtual or remote spaces in broadcast applications. Thus, the SMART-I^2 can add a high precision audio dimension to future 3D-TV systems by providing spatially coherent audio and visual rendering.

2. WAVE FIELD SYNTHESIS

2.1. Physical basis

Wave Field Synthesis (WFS) is a spatialized sound rendering technology which was first really developed at Delft University [3]. It is an audio implementation of Huygen's principle, which states that: *Every sound field emerging from one primary sound source can be reproduced by summing contributions of an infinite and continuous distribution of secondary sound sources.* At the theorical level, WFS allows one to synthesize a sound source at any given position. Implementations of WFS are simplified versions of this principle, typically using a linear array of equally spaced loudspeakers.

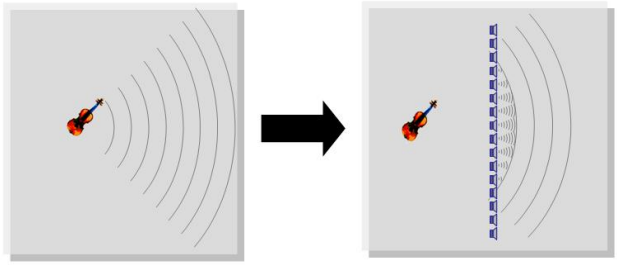

Fig. 1. Illustration of sound rendering using WFS.

Figure 1 illustrates the principle of WFS. The violin on the left part is the primary source producing the target natural sound field. The linear array of secondary sound sources on the right produces, through summation of the contributions of each loudspeaker driven appropriately, a synthesized sound

*marc.rebillat@limsi.fr

field equivalent to the original target field. The sound field of the virtual violin is synthesized, perceived by users in the reproduction area as emanating from the precise spatial location of the violin. Additional sound sources may be simultaneously synthesized through simple linear superposition.

2.2. Spatialized sound rendering

Using this physical basis, different types of fundamental sound sources, or sound fields, can be synthesized (see figure 2). Plane waves represent sound objects situated far away from the immersion area and are perceived as coming from a constant angle, independent of listener position. This can be used, for example, to render current spatial audio broadcasts (surround sound type) with ideal speaker placement for all users. Point sources represent sound objects near the immersion area. Point sources can be synthesized at positions behind or in front of the loudspeaker array. Such focused sources (*i.e.* point sources in front of the array) are perceived as being physically present in the immersion area.

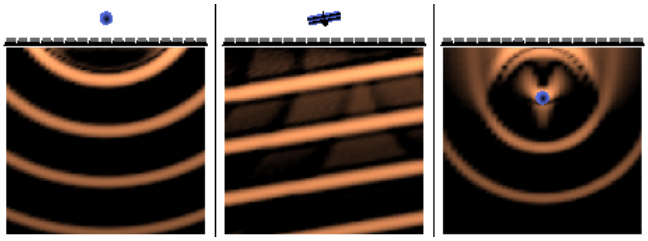

Fig. 2. Synthesis of different wave types. *Left:* Point source behind the loudspeaker array. *Center:* Plane wave. *Right:* Point source in front of the loudspeaker array.

Sound rendering using WFS, due to its approach in physically recreating the entire field within a spatially large area, is not limited to a single user at a single location. The sound perspective, including parallax, is correct for every user in the immersion area, without the need of a tracking device.

2.3. WFS in practice

As previously stated, practical WFS implementations are limited to a linear array and hence reproduction is optimized for the horizontal plane. Auditory perception is the most precise and more stable in the horizontal plane and therefore a more pertinent choice for array orientation. With this restriction, which reduces the required calculation power, the digital audio processing can be done with a latency of less than 5 ms. This is more than sufficient for real-time AV applications. Other limitations induced by the use of discrete and finite arrays are discussed in [5].

3. DESIGN OF THE SMART-I^2

This section provides a brief overview of the SMART-I^2 system. A more complete technical description of the system can be found in [6].

3.1. Tracked Stereoscopy

To produce a 3D visual rendering, each eye of the user must see the scene from a slightly different point of view. One means of realizing this is to use light polarization properties to independently address each eye of the user. The user wears special polarized glasses for visual cross-talk cancellation. The graphic rendering should also be adapted to the user's head position in order to maintain the correct point of view. Using this approach, the 3D visual rendering is coherent regardless of the user's position in the viewing area. This technique is referred as tracked passive stereoscopy (TPS).

3.2. Audio-visual integration with MAP

The integration of the two different technologies, TPS (see section 3.1) and WFS (see section 2), is achieved through an innovative use of multi-actuator panels (MAPs) [4]. MAPs are stiff lightweight panels with multiple electro-mechanical exciters attached to the backside. Typical MAP multichannel loudspeakers are not larger than 1 m^2. For this project, a novel large dimension MAP has been designed, (*i.e.* 5 m^2 with a $4/3$ ratio) in order to provide sufficient surface area and size to be used as a projection screen. To accommodate polarized light projection, the front face of the panel has been covered with metallic paint designed to preserve light polarization. Due to the nature of the MAP design, screen displacements caused by acoustic vibrations are very small and do not disturb 3D video projection on the surface of the panel. Such a structure then allows one to efficiently integrate a 3D visual rendering technology and a spatialized sound rendering technology.

3.3. Architecture

The hardware architecture of the SMART-I^2 is schematically presented in figure 3. Two large MAPs of 2.6 m $\times$ 2 m form a corner of stereoscopic screens and a 24 loudspeakers array. With this configuration, users can move within an immersion area of approximately 2.5 m $\times$ 2.5 m.

The rendering architecture of the SMART-I^2 is composed of three components. Virtual Choreographer (VirChor) [7] is an open source real-time 3D graphics engine that relies on an XML-based definition of 3D scenes with graphic and sonic components. The WFS Engine is a real-time audio engine dedicated to low latency WFS rendering, capable of real-time filtering of up to 24 input channels (virtual sources) routed to 24 output channels (every exciter of the two MAPs). Max/MSP [8] is a real-time audio analysis/synthesis engine

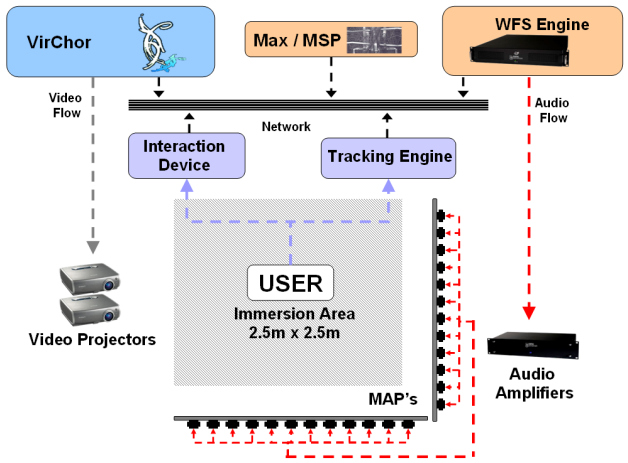

Fig. 3. Global overview of the SMART-I^2.

using a graphical programming environment which, in addition to sound synthesis, provides peripheral interfacing, timing, and communications. An example of a simple AV scene rendered by the SMART-I^2 is given in figure 4.

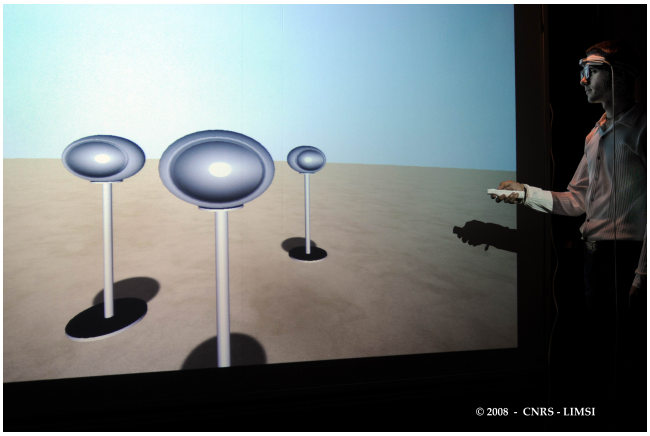

Fig. 4. An example of AV scenes provided the SMART-I^2.

4. PERCEPTUAL VALIDATION OF AZIMUTH PERCEPTION

Some preliminary perceptual tests have been performed in order to validate the usability of the SMART-I^2 with regards to AV 3D coherence. Fourteen subjects from 22 to 53 years old participated in the experiment. A complete description of the evaluation, as well as additional tests, are presented in [6].

4.1. Experimental protocol

Localization in the azimuth plane is the most basic means by which humans identify different sounds or decompose a complex sound scene. A test has been designed to evaluate the ability of users to locate in azimuth an AV source synthesized by the SMART-I^2. The results of this test aid in determining the degree to which AV coherence can be obtained without the use of semantic or contextual information.

For this test, subjects stood at the center of the immersion area and were instructed not to move. They faced an arc of 17 virtual loudspeakers, positioned every 3 $^\circ$ in azimuth and at a distance of 4 m from the center (behind the screen/array). After automatic confirmation of the subject's head position and orientation, a single 150 ms noise burst was played from one of the virtual loudspeakers. Subjects had to then indicate from which of the 17 objects the sound was coming. Only 7 of the loudspeakers were actually used as potential audio sources, with 15 repetitions of each of the 7 positions. The endmost virtual loudspeakers were not among the 7 actual source positions used.

4.2. Results

Results of these tests are shown in figure 5 using a statistical boxplot representation. Target sound source positions are given on the x-axis with corresponding subject responses on the y-axis. The ideal response line is shown as a reference. The boxes contains 50% of all subject answers. The whiskers (line extending from each end of the boxes) show the extreme values beyond the lower and the upper quartile. Outliers are indicated with red '+'.

What can be seen from these results is that subjects were able to locate a sound source with a median error of less than 3 $^\circ$, which is the angular separation between two neighboring speakers. For extremal target sources, the median error is close to -3 $^\circ$ for the leftmost, and 3 $^\circ$ for the rightmost. This indicates a tendency towards exaggeration of the angular positions of lateral sources. The half inter-quartile range (HIQR) (which can be interpreted as a standard deviation for a non Gaussian repartition) is always lower than 3 $^\circ$ except for the target source at 6 $^\circ$, for which the HIQR reached 4 $^\circ$.

4.3. Discussion

The absolute human localization blur, *i.e.* the accuracy with which a human can locate a real sound source, is between 1 $^\circ$ and 4 $^\circ$ in front of the listener [9], depending on the test signal and on the reporting method. The absolute localization bias of the present test is thus very close to natural human ability.

Measurements have been done in the entire immersion area. Using an auditory model, it is shown in [6] that audio localization cues quality is stable throughout. As such, the results shown in the previous test at the center position can be extrapolated as valid in the entire immersion area.

Moreover, in an AV context there can be additional benefit due to the ventriloquism effect, where there is some semantic correlation between audio and visual objects. If the lag

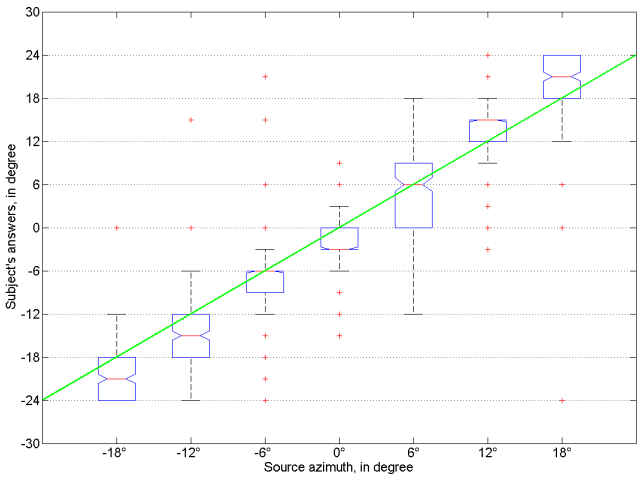

Fig. 5. Results of the azimuth perception test.

between the audio and visual stimuli is less than 100 ms, audio and visual stimuli are perceived as coming from the same azimuth if there are less than 3 ° between them [10]. In a situation where coherent AV objects are presented together, the ventriloquism effect will assure the perceptive fusion between the audio and visual stimuli. Consequently, the SMART-I^2 is able to provide the users with AV scenes perceptually coherent in azimuth.

5. POTENTIAL APPLICATIONS

In the 3D-TV community, great deal of efforts has been put recently into capturing and renderering 3D video for multiple users. The audio community is currently developing several performant methods to record and store 3D sound scenes. However, the combination of 3D audio and 3D video rederings remains a relatively unaddressed topic. The SMART-I^2 thus provides a first attempt for seamless real-time combination of high quality 3D audio and 3D video. In a broadcast application context, the SMART-I^2 is then able to render in real-time and in a spatially coherent way, any 3D live or recorded AV stream. Home theaters of the future could use such similar concepts.

Thanks to the large size of its rendering area, the SMART-I^2 can equally address multiple users. The system is currently limited to one user due to the visual rendering technology employed. This limitation could be removed through the use of combined tracked passive and active stereoscopy for multiple user experience. For broadcast installations, the combination of auto-stereoscopic displays or Parallax Panoramagrams with WFS audio rendering would allow for the rendering of the audio and visual parallax effects for multiple users simultaneously without the need for tracking systems or special glasses to be worn.

6. CONCLUSION

The SMART-I^2 is a successful means of combining demands of high quality spatialized audio rendering and 3D video rendering through an innovative use of multi-actuators panels.

A prototype has been build and evaluated. Perceptual experiments have been performed to validate the AV rendering in terms of azimuth localization accuracy. Experiments show that audio and visual stimuli provided by the SMART-I^2 can be perceptually integrated into one AV percept. The SMART-I^2 is thus a platform that can be used in a wide range of AV applications.

7. REFERENCES

[1] J. Blauert, *Communication Acoustics (Signals and Communication Technology)*, Springer-Verlag, 2005.

[2] W.P.J. de Bruijn, *Application of wave field synthesis in videoconferencing*, Ph.D. thesis, Delft University of Technology, 2004.

[3] A. J. Berkhout, D. de Vries, and P. Vogel, "Acoustic control by wave field synthesis," *Journal Of The Acoustical Society Of America*, vol. 93, no. 5, pp. 2764–2778, 1993.

[4] M. M. Boone, "Multi-actuator panels (MAP) as loudspeaker arrays for wave field synthesis," *Journal Of The Audio Engineering Society*, vol. 52, no. 7-8, pp. 712–723, 2004.

[5] E. Corteel, "Equalization in an extended area using multichannel inversion and wave field synthesis," *Journal Of The Audio Engineering Society*, vol. 54, no. 12, pp. 1140–1161, 2006.

[6] M. Rébillat, E. Corteel, and B.F.G. Katz, "SMART-I^2: Spatial multi-user audio-visual real-time interactive interface," *125th Convention of the Audio Engineering Society*, 2008.

[7] C. Jacquemin, "Architecture and experiments in networked 3d audio/graphic rendering with virtual choreographer.," in *Proceedings, Sounds and Music Computing (SMC'04), Paris.*, 2004.

[8] D. Zicarelli, G. Taylor, J.K. Clayton, R. Dudas, and B. Nevil, "Max 4.6: Reference manual, http://www.cylcing74.com," .

[9] J. Blauert, *Spatial Hearing, The Psychophysics of Human Sound Localization*, MIT Press, 1999.

[10] J. Lewald, W. H. Ehrenstein, and R. Guski, "Spatio-temporal constraints for auditory-visual integration," *Behavioural Brain Research*, vol. 121, no. 1-2, pp. 69–79, 2001.